# Ejercicio 2: Errores de Medición y Supuestos CEV

Errores en Variables Clásicos (Classical Errors-in-Variables)

## Contexto del Problema

Queremos estimar la siguiente ecuación:

$$\text{tvhours}^* = \beta_0 + \beta_1 \text{age} + \beta_2 \text{age}^2 + \beta_3 \text{motheduc} + \beta_4 \text{fatheduc} + \beta_5 \text{sibs} + u \tag{1}$$

donde:

- $\text{tvhours}^*$ = horas **verdaderas** de ver televisión por semana (variable **no observable**)
- tvhours = horas **reportadas** en la encuesta (variable **observable**)
- $u$ = error del modelo (término de perturbación estocástica)

**Problema fundamental**: No observamos $\text{tvhours}^*$, solo observamos tvhours.

### Modelo de error de medición

Asumimos que la relación entre el valor verdadero y el observado es:

$$\text{tvhours} = \text{tvhours}^* + e \tag{2}$$

donde $e$ es el **error de medición**.

## Pregunta a: ¿Se cumplen los supuestos CEV?

### Respuesta

Es **poco probable** que se cumplan perfectamente los supuestos de errores en variables clásicos (CEV) en esta aplicación. Analicemos cada supuesto:

**1. Media cero:** $E[e] = 0$

**Supuesto**: En promedio, no hay sesgo sistemático en el reporte.
**¿Se cumple?** PROBABLEMENTE NO
**Razones**:

- **Sesgo de deseabilidad social**: Ver mucha televisión puede percibirse negativamente. Los padres o niños pueden subreportar sistemáticamente las horas de TV para dar una mejor impresión.
- **Sesgo de memoria selectiva**: Las personas tienden a recordar mejor ciertos eventos que otros. Pueden olvidar horas "pasivas"de TV mientras hacen otras actividades.
- **Redondeo sistemático**: Si todos redondean hacia abajo (ej. ünas 2 horasçuando en realidad son 2.7), entonces $E[e] < 0$.

**Implicación formal**: Si $E[e] \neq 0$, entonces:

$$E[\text{tvhours}] = E[\text{tvhours}^*] + E[e] \neq E[\text{tvhours}^*] \tag{3}$$

El nivel promedio reportado difiere sistemáticamente del verdadero, pero esto **NO sesga los coeficientes de pendiente** si los demás supuestos se mantienen.

**2. Independencia del valor verdadero:** $\mathbf{Cov(tvhours^*}, e) = 0$

**Supuesto**: El error de medición no se correlaciona con el valor verdadero.
**¿Se cumple?** DUDOSO
**Razones por las que podría no cumplirse**:

1. **Heterogeneidad en la precisión del reporte**:
   - Niños que ven mucha TV (valores altos de $\text{tvhours}^*$) pueden perder la cuenta más fácilmente
   - El error podría aumentar con el nivel de TV: $\text{Var}(e|\text{tvhours}^*) = \sigma_e^2(\text{tvhours}^*)$
2. **Sesgo diferencial de reporte**:
   - Quien ve mucha TV puede sentir más presión para subreportar (sesgo social más fuerte)
   - Esto crearía: $E[e|\text{tvhours}^* = \text{alto}] < 0$ y $E[e|\text{tvhours}^* = \text{bajo}] \approx 0$
   - Por lo tanto: $\text{Cov}(\text{tvhours}^*, e) < 0$

**Implicación**: Si este supuesto falla, el error de medición deja de ser çlásicoz puede generar sesgos más complejos en los estimadores.

**3. Independencia de las X's: Cov$(X_j, e) = 0$ para toda $X_j$**

**Supuesto**: El error de medición no se correlaciona con las variables explicativas.
**¿Se cumple?** PROBABLEMENTE NO
**Razones específicas en este contexto**:

1. **Correlación con educación parental** (motheduc, fatheduc):
   - Padres con mayor educación pueden:
     - Ser más conscientes del estigma de ver mucha TV
     - Subreportar más las horas de TV de sus hijos
     - Tener mejor memoria/registros de actividades
   - Esto implica: Cov$(\text{motheduc}, e) < 0$
2. **Correlación con edad** (age):
   - Niños más pequeños: reportes hechos por padres (posiblemente más precisos)
   - Niños mayores: auto-reporte (posiblemente menos preciso o con más sesgo social)
   - Adolescentes pueden ser menos cooperativos en encuestas
   - Esto puede crear: Cov$(\text{age}, e) \neq 0$
3. **Correlación con número de hermanos** (sibs):
   - Familias grandes: más caótico, más difícil hacer seguimiento
   - Mayor error de medición en familias numerosas
   - Posible: Cov$(\text{sibs}, e) > 0$ en términos de varianza del error

**Implicación crítica**: Si Cov$(X_j, e) \neq 0$, entonces el error de medición se comporta como una variable omitida en la ecuación transformada, lo que puede sesgar los estimadores de todos los coeficientes.

### Conclusión sobre el cumplimiento de CEV

**Evaluación General**

Los supuestos CEV **probablemente NO se cumplen perfectamente** en esta aplicación debido a:

1. Sesgo social de deseabilidad $\rightarrow E[e] \neq 0$
2. Error diferencial según nivel de TV $\rightarrow \text{Cov}(\text{tvhours}^*, e) \neq 0$
3. Correlación con educación parental y edad $\rightarrow \text{Cov}(X_j, e) \neq 0$

Sin embargo, los supuestos CEV pueden ser una **aproximación razonable** si:

- Las violaciones son pequeñas en magnitud
- Los sesgos se cancelan parcialmente
- Las correlaciones son débiles

# Pregunta b: ¿Qué exigen los supuestos CEV en esta aplicación?

## Definición formal de los supuestos CEV

Los supuestos de **errores en variables clásicos** (Classical Errors-in-Variables) son:

**Supuesto 1** (Media Cero del Error de Medición)**.**

$$E[e] = 0 \tag{4}$$

**Interpretación en este contexto**:

- En promedio, los niños reportan exactamente sus horas verdaderas de TV
- Algunos sobre-reportan, otros sub-reportan, pero en promedio se cancela
- No hay sesgo sistemático hacia arriba o hacia abajo

**Ejemplo numérico**: Si 100 niños ven verdaderamente 10 horas/semana:

- Algunos reportarán 8, 9, 11, 12 horas
- El promedio de los reportes debe ser $\approx 10$ horas
- $\frac{1}{100}\sum_{i=1}^{100} e_i \approx 0$

**Supuesto 2** (Independencia del Valor Verdadero)**.**

$$Cov(tvhours^*, e) = 0 \quad o\ equivalentemente \quad E[e|tvhours^*] = 0 \tag{5}$$

**Interpretación en este contexto**:

- El error de medición no depende de cuánta TV realmente ve el niño
- Un niño que ve 5 horas tiene el mismo error (en expectativa) que uno que ve 20 horas
- La magnitud o dirección del error no está relacionada con el nivel verdadero de TV

**Ejemplo de violación**:

- Si niños que ven mucha TV (ej. 25 horas/semana) sistemáticamente subreportan más debido a vergüenza
- Entonces: $E[e|\text{tvhours}^* = 25] < 0$ y $E[e|\text{tvhours}^* = 5] \approx 0$
- Esto viola el supuesto

**Supuesto 3** (Independencia de las Variables Explicativas)**.**

$$Cov(X_j, e) = 0 \quad para\ j = 1, 2, \ldots, k \tag{6}$$

*o equivalentemente:*

$$E[e|age, age^2, motheduc, fatheduc, sibs] = 0 \tag{7}$$

**Interpretación específica en este contexto**:

1. **Cov**(**age**, $e$) = 0:
   - El error de medición no varía sistemáticamente con la edad del niño
   - Niños de 8 años cometen el mismo tipo de errores que niños de 14 años
2. **Cov**(**age**$^2$, $e$) = 0:
   - El error no tiene una relación cuadrática con la edad
3. **Cov**(**motheduc**, $e$) = 0 **y Cov**(**fatheduc**, $e$) = 0:
   - El error de medición no depende de la educación de los padres
   - Padres con educación universitaria reportan con la misma precisión que padres con educación secundaria
   - No hay diferencias en sesgo social relacionadas con educación
4. **Cov**(**sibs**, $e$) = 0:
   - El error no varía con el número de hermanos
   - Familias grandes y pequeñas reportan con igual precisión

**Supuesto 4** (Homocedasticidad del Error de Medición (opcional pero común))**.**

$$Var(e|X, tvhours^*) = \sigma_e^2 \tag{8}$$

**Interpretación**:

- La variabilidad del error es constante
- No depende de las características del niño ni del nivel verdadero de TV
- Todos tienen la misma "dispersión.en su error de reporte

## Implicaciones de los supuestos CEV

### Bajo los supuestos CEV, ¿qué ocurre con la estimación?

Cuando el error de medición está en la **variable dependiente** $Y$ (como en este caso), los supuestos CEV implican:

**Modelo verdadero**:

$$\text{tvhours}^* = X\beta + u \tag{9}$$

**Modelo estimado** (con error de medición):

$$\text{tvhours} = \text{tvhours}^* + e = X\beta + u + e = X\beta + v \tag{10}$$

donde $v = u + e$ es el nuevo término de error.

**Propiedades de** $v$:

1. **Media cero**:

$$E[v] = E[u + e] = E[u] + E[e] = 0 + 0 = 0 \quad \checkmark \tag{11}$$

2. **Independencia de X** (supuesto clave de Gauss-Markov):

$$\begin{aligned} E[v|X] &= E[u + e|X] && (12)\\ &= E[u|X] + E[e|X] \quad \text{(por supuesto 3 de CEV)} && (13)\\ &= 0 + E[e] \quad \text{(por supuesto 1 de CEV)} && (14)\\ &= 0 + 0 = 0 \quad \checkmark && (15) \end{aligned}$$

3. **Varianza aumentada**:

$$\begin{aligned} \text{Var}(v) &= \text{Var}(u + e) && (16)\\ &= \text{Var}(u) + \text{Var}(e) + 2\text{Cov}(u, e) && (17)\\ &= \sigma_u^2 + \sigma_e^2 + 0 \quad \text{(si } u \text{ y } e \text{ son independientes)} && (18)\\ &= \sigma_u^2 + \sigma_e^2 > \sigma_u^2 && (19) \end{aligned}$$

Resultado Principal

**Conclusión**: Bajo los supuestos CEV, cuando el error de medición está en $Y$:

1. Los estimadores MCO de $\beta$ son **INSESGADOS**:

$$E[\hat{\beta}] = \beta \tag{20}$$

2. Pero tienen **MAYOR VARIANZA**:

$$\text{Var}(\hat{\beta}) = \sigma_v^2(X'X)^{-1} = (\sigma_u^2 + \sigma_e^2)(X'X)^{-1} > \sigma_u^2(X'X)^{-1} \tag{21}$$

3. Los errores estándar estimados son **más grandes**
4. Los tests de hipótesis tienen **menor potencia**
5. Los intervalos de confianza son **más amplios**

**En resumen**: Perdemos **eficiencia** pero no **insesgamiento**.

# Pregunta c: Derivación formal del sesgo por error de medición

## Setup del problema

**Modelo poblacional verdadero**:

$$\text{tvhours}^* = \beta_0 + \beta_1\text{age} + \beta_2\text{age}^2 + \beta_3\text{motheduc} + \beta_4\text{fatheduc} + \beta_5\text{sibs} + u \tag{22}$$

**Modelo de error de medición**:

$$\text{tvhours} = \text{tvhours}^* + e \tag{23}$$

**Modelo que estimamos en la práctica**:

$$\text{tvhours} = \beta_0 + \beta_1\text{age} + \beta_2\text{age}^2 + \beta_3\text{motheduc} + \beta_4\text{fatheduc} + \beta_5\text{sibs} + \varepsilon \tag{24}$$

donde $\varepsilon$ es el término de error en la regresión observada.

## Paso 1: Relacionar el modelo observado con el verdadero

Sustituyendo la ecuación (7) en (8):

$$\begin{aligned}
\text{tvhours} &= \text{tvhours}^* + e && (25)\\
&= (\beta_0 + \beta_1\text{age} + \beta_2\text{age}^2 + \beta_3\text{motheduc} + \beta_4\text{fatheduc} + \beta_5\text{sibs} + u) + e && (26)\\
&= \beta_0 + \beta_1\text{age} + \beta_2\text{age}^2 + \beta_3\text{motheduc} + \beta_4\text{fatheduc} + \beta_5\text{sibs} + (u + e) && (27)
\end{aligned}$$

Comparando con la ecuación (9), vemos que:

$$\varepsilon = u + e \tag{28}$$

**Conclusión del Paso 1**: El término de error en la regresión observada es la suma del error del modelo verdadero ($u$) y el error de medición ($e$).

## Paso 2: Analizar las propiedades del nuevo error

Para que los estimadores MCO sean insesgados, necesitamos:

$$E[\varepsilon|X] = 0 \tag{29}$$

donde $X$ representa todas las variables explicativas: $(\text{age}, \text{age}^2, \text{motheduc}, \text{fatheduc}, \text{sibs})$.
**Verificación**:

$$\begin{aligned} E[\varepsilon|X] &= E[u + e|X] && (30)\\ &= E[u|X] + E[e|X] && (31)\\ &= 0 + E[e|X] \quad \text{(por supuesto de exogeneidad del modelo verdadero)} && (32) \end{aligned}$$

Ahora, por el **Supuesto 3 de CEV** ($\text{Cov}(X_j, e) = 0$ para toda $X_j$):

$$E[e|X] = E[e] = 0 \quad \text{(por Supuesto 1 de CEV)} \tag{33}$$

Por lo tanto:

$$E[\varepsilon|X] = 0 + 0 = 0 \quad \checkmark \tag{34}$$

Resultado Importante

**Conclusión del Paso 2**:
El supuesto de exogeneidad **se mantiene** bajo los supuestos CEV. Esto implica que:

$$\boxed{E[\hat{\beta}] = \beta \quad \text{(Los estimadores son INSESGADOS)}} \tag{35}$$

**No hay sesgo en los coeficientes** cuando el error de medición está en la variable dependiente.

## Paso 3: Analizar la varianza de los estimadores

Aunque no hay sesgo, la varianza de los estimadores sí se ve afectada.
**Fórmula de la varianza del estimador MCO**:

$$\text{Var}(\hat{\beta}|X) = \sigma_\varepsilon^2 (X'X)^{-1} \tag{36}$$

donde $\sigma_\varepsilon^2 = \text{Var}(\varepsilon) = \text{Var}(u + e)$.
**Cálculo de $\sigma_\varepsilon^2$**:

Asumiendo que $u$ y $e$ son independientes (una extensión razonable de los supuestos CEV):

$$\begin{aligned}\sigma_\varepsilon^2 &= \text{Var}(u+e) && (37)\\ &= \text{Var}(u) + \text{Var}(e) + 2\text{Cov}(u,e) && (38)\\ &= \sigma_u^2 + \sigma_e^2 + 0 && (39)\\ &= \sigma_u^2 + \sigma_e^2 && (40)\end{aligned}$$

**Comparación**:

- **Varianza sin error de medición**:

$$\text{Var}(\hat{\beta}_{\text{true}}) = \sigma_u^2 (X'X)^{-1} \tag{41}$$

- **Varianza con error de medición**:

$$\text{Var}(\hat{\beta}_{\text{obs}}) = (\sigma_u^2 + \sigma_e^2)(X'X)^{-1} \tag{42}$$

Por lo tanto:

$$\boxed{\text{Var}(\hat{\beta}_{\text{obs}}) > \text{Var}(\hat{\beta}_{\text{true}})} \tag{43}$$

**El error de medición infla la varianza de los estimadores**.

## Paso 4: Implicaciones para inferencia estadística

### 4.1. Errores estándar

Los errores estándar estimados serán:

$$\widehat{\text{SE}}(\hat{\beta}) = \sqrt{\hat{\sigma}_\varepsilon^2 (X'X)^{-1}} \tag{44}$$

donde:

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n-k-1}\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \frac{1}{n-k-1}\sum_{i=1}^{n} (u_i + e_i)^2 \tag{45}$$

Como $\hat{\sigma}_\varepsilon^2$ estima $\sigma_u^2 + \sigma_e^2 > \sigma_u^2$:

$$\boxed{\widehat{\text{SE}}(\hat{\beta}) > \text{SE}(\hat{\beta}_{\text{true}})} \tag{46}$$

**Los errores estándar son más grandes**.

### 4.2. Estadísticos t

El estadístico t para probar $H_0 : \beta_j = 0$ es:

$$t_j = \frac{\hat{\beta}_j}{\widehat{\mathrm{SE}}(\hat{\beta}_j)} \tag{47}$$

Dado que:

- $E[\hat{\beta}_j] = \beta_j$ (numerador correcto)
- $\widehat{\mathrm{SE}}(\hat{\beta}_j)$ es más grande (denominador inflado)

Entonces:

$$|t_j| = \frac{|\hat{\beta}_j|}{\widehat{\mathrm{SE}}(\hat{\beta}_j)} \quad \text{tiende a ser } \textbf{más pequeño} \tag{48}$$

**Implicación**: Mayor probabilidad de **no rechazar** $H_0$ (error Tipo II), es decir, menor **potencia** del test.

### 4.3. Intervalos de confianza

Un intervalo de confianza al 95 % es:

$$\mathrm{IC}_{95\,\%}(\beta_j) = \left[\hat{\beta}_j - 1{,}96 \times \widehat{\mathrm{SE}}(\hat{\beta}_j),\ \hat{\beta}_j + 1{,}96 \times \widehat{\mathrm{SE}}(\hat{\beta}_j)\right] \tag{49}$$

Como $\widehat{\mathrm{SE}}(\hat{\beta}_j)$ es más grande:

$$\boxed{\text{Los intervalos de confianza son MÁS AMPLIOS}} \tag{50}$$

**Menor precisión en la estimación**.

### 4.4. $R^2$ (Bondad de ajuste)

El $R^2$ se calcula como:

$$R^2 = 1 - \frac{\sum \hat{\varepsilon}_i^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{SSR}{SST} \tag{51}$$

Con error de medición:

$$SSR = \sum (u_i + e_i)^2 > \sum u_i^2 \tag{52}$$

Por lo tanto:

$$\boxed{R^2_{\text{obs}} < R^2_{\text{true}}} \tag{53}$$

**El ajuste del modelo aparenta ser peor**.

## Demostración formal del sesgo: Caso general

Ahora generalizemos para entender qué pasa cuando el error está en una variable explicativa.

**Modelo verdadero**:

$$Y = \beta_0 + \beta_1 X^* + u \tag{54}$$

**Error de medición en $X$**:

$$X = X^* + e \tag{55}$$

**Modelo estimado**:

$$Y = \beta_0 + \beta_1 X + v \tag{56}$$

### Derivación del sesgo

Sustituyendo (31) en (30):

$$\begin{aligned} Y &= \beta_0 + \beta_1 X^* + u && (57)\\ &= \beta_0 + \beta_1 (X - e) + u && (58)\\ &= \beta_0 + \beta_1 X - \beta_1 e + u && (59)\\ &= \beta_0 + \beta_1 X + (u - \beta_1 e) && (60) \end{aligned}$$

Comparando con (32): $v = u - \beta_1 e$

**Verificar exogeneidad**:

$$\begin{aligned} \text{Cov}(X, v) &= \text{Cov}(X, u - \beta_1 e) && (61)\\ &= \text{Cov}(X, u) - \beta_1 \text{Cov}(X, e) && (62)\\ &= 0 - \beta_1 \text{Cov}(X^* + e, e) \quad \text{(si } u \text{ exógeno)} && (63)\\ &= -\beta_1 [\text{Cov}(X^*, e) + \text{Var}(e)] && (64)\\ &= -\beta_1 [0 + \sigma_e^2] \quad \text{(por Supuesto 2 de CEV)} && (65)\\ &= -\beta_1 \sigma_e^2 \neq 0 && (66) \end{aligned}$$

**¡El supuesto de exogeneidad se viola!**

### Límite probabilístico del estimador

El estimador MCO converge a:

$$\text{plim}(\hat{\beta}_1) = \beta_1 + \frac{\text{Cov}(X, v)}{\text{Var}(X)} \tag{67}$$

$$= \beta_1 + \frac{-\beta_1 \sigma_e^2}{\text{Var}(X^* + e)} \tag{68}$$

$$= \beta_1 + \frac{-\beta_1 \sigma_e^2}{\text{Var}(X^*) + \sigma_e^2} \quad \text{(por independencia de } X^* \text{ y } e) \tag{69}$$

$$= \beta_1 \left(1 - \frac{\sigma_e^2}{\text{Var}(X^*) + \sigma_e^2}\right) \tag{70}$$

$$= \beta_1 \left(\frac{\text{Var}(X^*)}{\text{Var}(X^*) + \sigma_e^2}\right) \tag{71}$$

Definiendo el **ratio de confiabilidad**:

$$\lambda = \frac{\text{Var}(X^*)}{\text{Var}(X^*) + \sigma_e^2} = \frac{\text{Var}(X^*)}{\text{Var}(X)} \in (0, 1) \tag{72}$$

Entonces:

$$\boxed{\text{plim}(\hat{\beta}_1) = \lambda \beta_1} \tag{73}$$

**Interpretación**:

- Como $0 < \lambda < 1$, el estimador está **atenuado** (sesgado hacia cero)
- Cuanto mayor sea $\sigma_e^2$ (más error de medición), menor será $\lambda$ y mayor el sesgo
- Esto se conoce como **sesgo de atenuación** (attenuation bias)

## Resumen de resultados

| **Situación** | **Error en Y** | **Error en X** |
|---|---|---|
| Sesgo en $\hat{\beta}$ | NO | SÍ (atenuación) |
| Varianza de $\hat{\beta}$ | Aumenta | Aumenta |
| Errores estándar | Más grandes | Más grandes |
| $R^2$ | Disminuye | Disminuye |
| Potencia de tests | Disminuye | Disminuye |

Cuadro 1: Efectos del error de medición según su ubicación

## Aplicación al ejercicio: tvhours

**En nuestro caso**:

- El error de medición está en `tvhours` (variable dependiente $Y$)

- Las variables explicativas (age, motheduc, etc.) se asumen medidas sin error

**Conclusión**:

$$\boxed{\text{NO hay sesgo en los coeficientes } \hat{\beta}_j} \tag{74}$$

**Pero**:

1. Los estimadores tienen mayor varianza: $\text{Var}(\hat{\beta}) = (\sigma_u^2 + \sigma_e^2)(X'X)^{-1}$
2. Los errores estándar son más grandes
3. Los tests son menos potentes (mayor probabilidad de error Tipo II)
4. Los intervalos de confianza son más amplios
5. El $R^2$ es más bajo

**Mensaje final**: El error de medición en $Y$ no destruye la validez de la inferencia causal (los coeficientes son insesgados), pero reduce la **eficiencia estadística** (precisión de las estimaciones).